INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Phone Classification using a Non-Linear Manifold with Broad Phone Class Dependent DNNs

Linxue Bai, Peter Jančovič, Martin Russell, Philip Weber, Steve Houghton

Department of Electronic Electrical and Systems Engineering,
The University of Birmingham, Birmingham B15 2TT, UK
{lxb190, p.jancovic, m.j.russell, s.houghton}@bham.ac.uk, dr.philip.weber@ieee.org

Abstract

Most state-of-the-art automatic speech recognition (ASR) systems use a single deep neural network (DNN) to map the acoustic space to the decision space. However, different phonetic classes employ different production mechanisms and are best described by different types of features. Hence it may be advantageous to replace this single DNN with several phone class dependent DNNs. The appropriate mathematical formalism for this is a manifold. This paper assesses the use of a non-linear manifold structure with multiple DNNs for phone classification. The system has two levels. The first comprises a set of broad phone class (BPC) dependent DNN-based mappings and the second level is a fusion network. Various ways of designing and training the networks in both levels are assessed, including varying the size of hidden layers, the use of the bottleneck or softmax outputs as input to the fusion network, and the use of different broad class definitions. Phone classification experiments are performed on TIMIT. The results show that using the BPC-dependent DNNs provides small but significant improvements in phone classification accuracy relative to a single global DNN. The paper concludes with visualisations of the structures learned by the local and global DNNs and discussion of their interpretations.

Index Terms: manifold, phone classification, neural network, non-linear mapping, fusion, speech

1. Introduction

Deep neural networks (DNNs) trained with phone posterior probability targets can be used to create very low-dimensional discriminative representations of speech, called bottleneck features (BNFs). In automatic speech recognition (ASR) experiments, BNFs with as few as 9 dimensions perform as well as 39 dimensional features based on conventional mel frequency cepstral coefficients (MFCCs) [1], have an intuitive dynamical structure, and can be interpreted in terms of human perception and production [2].

The non-linear function T realised by the DNN maps the "acoustic space" A (for example, the space of short sequences of vectors of log filter-bank energies) to the BNF space B (*en route* to the space of vectors of phone posterior probabilities). Although this is a single continuous mapping, in practice the DNN is trained to approximate a discontinuous function whose outputs jump between 0 and 1 across triphone state boundaries. Therefore, assuming that acoustic space A is connected (which we don't actually know), it may be advantageous to think of T as a set of continuous functions $\{T_1, ..., T_N\}$, where each T_i is defined on a subset $A_i \subseteq A$ and $A = \bigcup_i A_i$. In this case the appropriate mathematical structure is a non-linear topological manifold. Intuitively, one might hope that the subsets A_i correspond to broad phone classes (BPCs), so that the mappings T_i implement phone-class dependent feature extraction.

The idea of phone-dependent feature extraction is well-established. For example, while vocal tract resonance frequencies provide a natural description of vowels, unvoiced consonants are better described in terms of duration and mean energies in key frequency bands [3, 4, 5, 6, 7, 8, 9]. There are also a number of studies that use BPC-dependent classifiers to focus on subtle differences between phones within a BPC [10].

A two-level linear computational model that is motivated by these considerations is presented in [11]. The first level comprises a set of discriminative linear transforms W_j^T, one for each of a set of overlapping BPCs Q_j, $j = 1, ..., N$, that are used for feature extraction. The transforms W_j^T are obtained using variants of linear discriminant analysis (LDA). An acoustic feature vector t is transformed using each W_j^T to obtain $t_j = W_j^T t$ and k-nearest neighbour methods are used to estimate $p(Q_j|t_j)$ and $p(c|Q_j, t_j)$ for each specific phone class c. These probabilities are combined in the second level to estimate the posterior probabilities $p(c|t_j)$ and hence to classify t. In acoustic feature vector phone classification experiments on TIMIT [12], the two-level linear classifier obtained slightly better results when BPC-specific linear transforms were learned, compared to a single transform. The authors of [11] speculate that better performance would be achieved using non-linear DNN-based transformations.

This paper extends our previous study of very low-dimensional BNFs, including phone classification [1] and visualization and interpretation [2]. Our objective is to determine whether it is advantageous for phone-classification of feature vectors to treat the acoustic space A as a non-linear manifold, in which several BPC-dependent DNNs rather than a single DNN are used for phone classification. We use the phone classes from [11]. For a broad class Q_j ($j = 1, ..., N$), containing K_j phones, we train a DNN D_j to map an element $a \in A$ onto a $K_j + 1$ dimensional vector P_j of posterior probabilities, where $P_j(i)$ is the probability of phone i given a ($i = 1, ..., K_j$) and $P_j(K_j + 1)$ is the probability that a corresponds to a phone outside class Q_j. The final hidden layer of each DNN is a 9 dimensional bottleneck. Classification is achieved by applying a fusion network, either to the set of posterior probability vectors P_j or to the outputs of the bottleneck layers. The output of the fusion network is a vector of 49 phone posterior probabilities.

Various ways of designing and training the DNNs are assessed, including varying the size of bottleneck and intermediate hidden layers, the use of the bottleneck outputs or softmax outputs as input to the fusion network, and the use of different BPC definitions. The results show a small but significant improvement compared with a single DNN, when a non-linear manifold structure incorporating multiple BPC-

http://dx.doi.org/10.21437/Interspeech.2017-1179

dependent DNNs is used for phone classification.

2. Topological manifolds

In mathematics an n-dimensional manifold is a topological space that is locally equivalent to n dimensional real Euclidean space $\mathbb{R}^n$ (for example, [13]). A simple example of a 1-dimensional manifold is a circle C in the plane, because any point on C has a neighbourhood that can be "straightened out" to be an open interval in $\mathbb{R} = \mathbb{R}^1$. However note that C cannot be embedded as a whole as a subset of $\mathbb{R}^1$.

Formally, a manifold consists of a topological space M such that for each $x \in M$ there is a neighbourhood U_x and bijection $\phi_x : U_x \rightarrow \mathbb{R}^n$ that establishes the equivalence between U_x and $\mathbb{R}^n$. Normally additional restrictions are placed on ϕ_x to ensure that it preserves relevant mathematical structure. Thus, in topology ϕ_x would be a homeomorphism (ϕ_x and ϕ_x^{-1} are both continuous) but for the purposes of calculus it would need to be a diffeomorphism (ϕ_x and ϕ_x^{-1} are also both differentiable). There is also a "consistency" property. If $x, y \in M$ and $U_x \cap U_y \neq \emptyset$ then $\phi_x \phi_y^{-1} : \phi_y(U_x \cap U_y) \rightarrow \phi_x(U_x \cap U_y)$ is a bijection with the same additional properties as ϕ_x and ϕ_y that ensures that the overlap $U_x \cap U_y$ is treated equivalently by ϕ_x and ϕ_y.

In speech processing there are a number of computational models where an acoustic space M is embedded into $\mathbb{R}^n$ for some n by a single global mapping ϕ. For example, in speaker or language identification M is the set of sequences of acoustic vectors corresponding to a spoken utterance, $\phi : M \rightarrow \mathbb{R}^n$ maps $x \in M$ to its i-vector $\phi(x)$, or M is the set of acoustic feature vectors in context, $\phi : M \rightarrow \mathbb{R}^n$ is implemented by a DNN and $\phi(x)$ is a bottleneck feature representation of $x \in M$. In contrast, the linear model described in [11] is one of few examples which attempt to exploit the varying local structure offered by a manifold.

3. Phone classification system based on a non-linear manifold with neural networks

The proposed phone classification system is inspired by a non-linear manifold model of speech. Its structure (Figure 1), comprises two levels. The first level is a set of N parallel non-linear local mapping functions $\phi_i (i = 1, ..., N)$, each focusing on a particular part of the speech acoustic space A. The second level integrates the outputs from the individual local mappings in the first level to arrive at a final classification decision. The following subsections describe each level in the system.

3.1. A non-linear manifold using broad phone class DNNs

Speech sounds in different BPCs result from different articulatory mechanisms and lend themselves to different types of parameterizations. Hence we assume that these BPCs dictate the manifold structure of A. In the first level of the system, each ϕ_i is realised using a BPC-dependent DNN. These networks operate in parallel, with each local network defined on the whole of A but focusing on a particular BPC.

An input to each local network ϕ_i is an element of A comprising logarithm filter-bank energies (logFBEs) in context. All the training data is passed to each network, regardless of which broad class they belong to. This ensures that a given local network has information about data which do not belong to its BPC. Suppose that the i^{th} local DNN implements a mapping ϕ_i for the subset Q_i of phones in the i^{th} BPC. The output layer

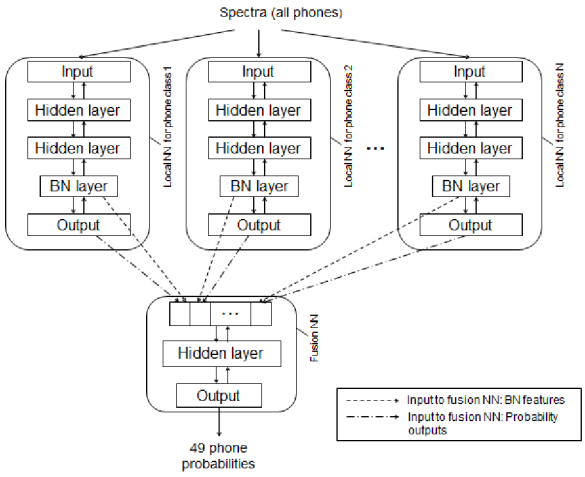


Figure 1: *An architecture of the phone classification system exploiting DNN-based manifold learning of speech.*

of this network has $K_i + 1$ nodes, where the first K_i nodes correspond to each phone in the category Q_i and the additional node is used to indicate the 'out-of-the-class' label used for input features which are not contained in the i^{th} BPC. The targets in the output layer are the phone or 'out-of-the-class' posterior probabilities. A nine-dimensional bottleneck layer is kept in each broad phone class network to enable comparison with the baseline single global bottleneck neural network.

We explored the use of different definitions of BPCs. As the base, the phones are grouped into the 8 non-overlapping BPCs from [11]. These correspond to BPCs Q_1–Q_8 in the upper part of Table 1. We also use several 'super' broad classes which are unions of two or more BPCs from Q_1–Q_8. These are defined in the lower half of Table 1. The model comprising A and the set of mappings ϕ_i from A into the 9-dimensional BNF feature space B, is inspired by the manifold framework described in Section 2. However, it falls short of the formal definition of a topological manifold, since each of the BPC-dependent DNNs is defined on the complete acoustic space A, rather than a subset corresponding to that BPC, and there is no guarantee that the mappings ϕ_i are homeomorphisms.

In order to eliminate the effect of unclear and ambiguous boundaries in TIMIT labels, we also explored neural networks that are re-trained and fine-tuned to features from centre frames of the phones.

3.2. Fusing the manifold information

The second level of the classification system is a fusion network that serves to provide the final phone classification decision. The input vector passed to this second level contains information concatenated from all the first level broad phone class DNNs. This could be the outputs of the bottleneck (or a hidden) layer, or the softmax probability outputs from the output layer of each first level networks. This is indicated by dashed and dot-dashed lines in Figure 1. The output of the fusion network is a vector of posterior phone probabilities of 49 phones.

Table 1: *Phonetic broad classes used to define the set of local DNN-based projections.*

Group	Phonetic class	Phone label
Q_1	Plosive	/g/, /d/, /b/, /k/, /t/, /p/
Q_2	Strong fricative	/s/, /z/, /sh/, /zh/, /ch/, /jh/
Q_3	Weak fricative	/f/, /v/, /th/, /dh/, /hh/
Q_4	Nasal/Flap	/m/, /n/, /en/, /ng/, /dx/
Q_5	Semi-vowel	/l/, /el/, /r/, /w/, /y/
Q_6	Short vowel	/ih/, /ix/, /ae/, /ah/, /ax/, /eh/, /uh/, /aa/
Q_7	Long vowel	/iy/, /uw/, /ao/, /er/, /ey/, /ay/, /oy/, /aw/, /ow/
Q_8	Silence	/sil/, /epi/, /q/, /vcl/, /cl/
Q_9	$Q_5 \cup Q_6 \cup Q_7$:	Semi-vowel, Short vowel, Long vowel
Q_{10}	$Q_1 \cup Q_3$:	Plosive, weak fricative
Q_{11}	$Q_5 \cup Q_6$:	Semi vowel, Short vowel
Q_{12}	$Q_5 \cup Q_7$:	Semi vowel, Long vowel
Q_{13}	$Q_6 \cup Q_7$:	Short vowel, Long vowel
Q_{14}	$Q_1 \cup Q_2 \cup \ldots \cup Q_8$:	All phones

4. Experimental setup

4.1. Methodology

Experiments were performed on the TIMIT speech corpus [12] whose training set contains recordings of 462 speakers, having in total 3696 utterances of 142910 tokens. We used 90% of the training set utterances, selected randomly for each gender in each dialect, as the neural network training set, and the remaining 10% as validation set. The SA recordings were excluded in order to avoid any bias due to their identical content. The 61 phone set used in TIMIT labels is mapped to the 49 phone set [14], which was used as targets for training.

The systems were evaluated with respect to their ability of classifying phones at the feature vector level. Two sets of experimental evaluations were performed: i) using all the feature vectors, and ii) using only the centre feature vector from each TIMIT phone segment. For evaluating phone classification accuracy, the 49 phone set was reduced to 40 according to [14]. The reported results are based on the core test set, which contains speech from 24 speakers.

4.2. Parameter setup

4.2.1. Speech feature representation

The speech signal, sampled at 16 kHz, was analysed using a 25 ms Hamming window with a 10 ms frame rate. We used a linear-frequency filter-bank analysis, with the number of filter-bank channels set to 26. The resulting logFBEs were normalised to have zero mean and unit variance based on the entire training set. The feature vector used as input to the neural network contained the logFBEs of the current frame and five preceding and five following frames, resulting in a 286 dimensional vector.

4.2.2. Training of global and local broad phone class neural networks

The networks were trained as deep belief networks (DBNs) with Gaussian-binary restricted Boltzmann machines(GRBMs) and restricted Boltzmann machines (RBMs) pre-training and stochastic gradient decent back propagation using the Theano toolkit [15, 16]. The learning rates of GRBM, RBM and fine-tuning were 0.002, 0.02, 0.1 respectively. In the fine-tuning process the training stopped when the error on the validation set started to rise or when the epoch reached the maximum of 100. For the experiments on centre feature vectors, the neural networks were further fine-tuned using only feature vectors corresponding to the centres of the TIMIT phone segments.

4.3. DNN structures

The structures of the different DNNs employed in the system were determined empirically. For the single global network a DNN of type $286-1024-1024-9-49$ was used, as this was found to achieve best performance. In the two-level networks, the BPC-dependent DNN for class Q_i in the first level is of type $256-256-9-(K_i+1)$. The fusion network is $X-32-49$, where $X = N + \sum_{i=1}^{N} K_i$, where N is the number of BPCs.

5. Results and discussion

5.1. Phone classification results

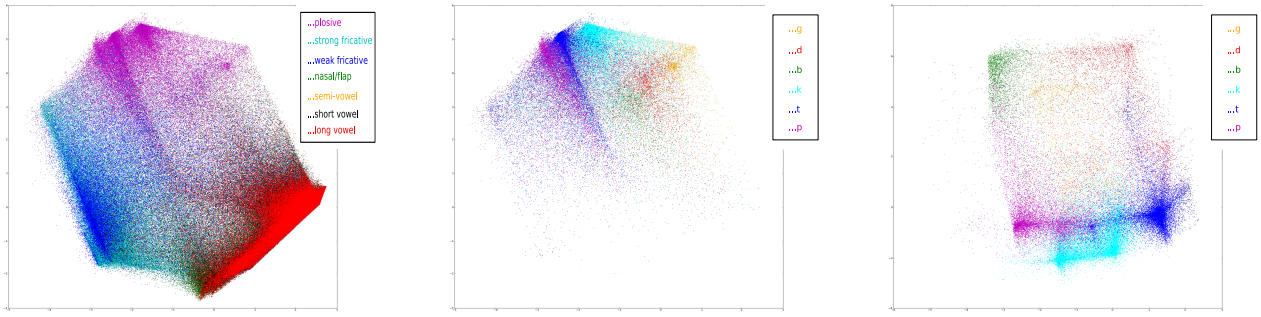
We considered five different sets of BPCs, D_1 to D_5 (Table 2). These determine the sets of local BPC-dependent DNNs. D_1 consists of the 8 non-overlapping broad phone classes Q_1 to Q_8 (Table 1). D_2 consists of D_1 plus the additional class Q_9 which combines the three vowel sub-categories. D_3 consists of D_2 plus Q_{10}, which combines plosives with weak fricatives. D_5 is the same as that used in [11], with additional classes (Q_{11}, Q_{12} and Q_{13}) for different combinations of vowel sub-categories plus a global class Q_{14} containing all phones. D_4 is D_5, without the global class Q_{14}.

Table 2: *The sets $D_1, ..., D_5$ of BPCs used to train local BPC-dependent DNNs in the two-level system.*

Broad phone class	Experimental setup				
	D_1	D_2	D_3	D_4	D_5
$Q_1 - Q_8$	X	X	X	X	X
Q_9		X	X		
Q_{10}			X	X	X
Q_{11}				X	X
Q_{12}				X	X
Q_{13}				X	X
Q_{14}					X
# of local DNNs	8	9	10	12	13

Tables 3 shows classification results for all frames and centre frames only for the single global DNN and the two-level manifold structures corresponding to $D_1, ..., D_5$. The figures are the averages of experiments performed over 20, 10 and 6 random DNN parameter initialisations for the global, local (softmax) and local (BNF), respectively. The single global DNN has approximately 2.44 million parameters, compared with 1.14, 1.28, 1.42, 1.71, and 1.85 million parameters for the two-level systems based on D_1 to D_5, respectively.

Focussing first on classification using all frame feature vectors, the average phone recognition accuracy for the single global DNN is 67.60% with standard deviation of 0.48. The two-level structure with local DNNs gives in all cases better performance, which in many cases presents a significant improvement ("*" indicates that in 95% of pairwise comparisons between global and local networks, the difference in performance is significant at the 0.05 level according to the McNe-



(a) (b) (c)

Figure 2: *Visualisations of LDA-based projections of 9-dimensional bottleneck features from Q_1 ('plosive') broad phone class DNN. 1^{st} vs. 2^{nd} dimension for all data (a); 1^{st} vs. 2^{nd} dimension (b) and 3^{rd} vs. 4^{th} dimension (c) for data within plosive class only.*

mar's test [17]). The two-level systems corresponding to D_2 to D_5 give good improvement over D_1 setup. The best performance (70.01% accuracy) is obtained with D_5. These results indicates that it is useful to include local network(s) that operate on a union of two or more BPCs, in particular, the union of vowel sub-categories or combination of pairs of vowel sub-categories, and also plosives with weak fricatives. This reflects the similarity in production of these sub-categories. The inclusion of such super-broad class networks may help to account for errors due to possible confusion between broad phone categories. For instance, the results in [11] indicate that there was a considerable confusion between these categories when performing classification of broad phone classes.

The results using only the centre feature vectors of each phone segment show similar performance trends to those observed for all feature vectors, however accuracy is considerably higher. This indicates that the classification error is higher during phone transitions. Fusing the outputs of the BNF layer gives slightly poorer performance than the softmax layer.

Table 3: *Phone classification accuracy obtained using all signal frames and using only the centre frames of each phone.*

	All frames		Centre frames	
Global DNN	67.60 (avg)		76.81 (avg)	
	69.05 (avg+3std)		77.58 (avg+3std)	
Local DNNs	Fusion net input		Fusion net input	
	Softmax	BNF	Softmax	BNF
D_1 (avg)	69.05*	68.78	77.45	77.03
D_2 (avg)	69.44*	69.23*	77.85	77.75
D_3 (avg)	69.56*	69.24*	78.31*	78.11*
D_4 (avg)	69.76*	69.31*	78.59*	78.08
D_5 (avg)	70.01*	69.63*	78.93*	78.70*

5.2. Visualisations of bottleneck features from local DNNs

This section explores visualisations of the structures learned by local BPC-based DNNs. The 9 dimensional bottleneck features from the local DNNs are projected onto 2D space using linear discriminant analysis (LDA). Example 2D visualisations for BPC Q_1 (plosives) are shown in Figure 2.

Figure 2(a) shows the first 2 dimensions for data from all phones (excluding silence for clarity). Plosives are represented in purple. Interestingly, although the non-plosive data were all assigned to 'out-of-the-class' category, the network has structured this data in an unsupervised manner according to phonetic categories. Figure 2(b) shows data for 'plosive' phones from Figure 2(a), with each plosive represented in a different colour. It can be seen that $/p/$, $/t/$, and $/k/$ are placed in an order which reflects their place of articulation (lips, teeth and soft palate, respectively). The voiced counterparts $/b/$, $/d/$, $/g/$ are placed in the same order but shifted towards the lower right. Figure 2(c) shows the plosive phone data projected onto LDA dimensions 3 and 4. Again, good separation of each plosive class is evident. Dimension 4 now seems to indicate voicing, with the unvoiced plosives placed in the lower part and voiced in the higher part of the figure. Again, the location structure for the unvoiced plosives is the same as for the voiced plosives, but shifted in dimension 4 for voicing.

6. Summary and future work

This paper presented phone classification system inspired by a non-linear manifold model of speech acoustic space. This system comprised of two levels. The first level consisted of a set of broad phone class (BPC) dependent DNNs. A representation from the output or hidden bottleneck layer of the first level network was used as input to the second level fusion network. Experimental evaluations were performed on TIMIT. The results showed that using the BPC-dependent DNNs provided small but significant improvements in phone classification accuracy in comparison to a single global DNN. It was demonstrated that in addition to the use of a set of local DNNs corresponding to basic BPCs, it was advantageous to also include local DNNs focusing on a combination of some BPCs, especially, vowel sub-categories. The use of the bottleneck or softmax outputs as input ot the fusion network provided similar results. Visualisations of the structures learned by the local DNNs indicated a relationship to speech production mechanisms.

To obtain a true topological manifold, the 'local' non-linear mappings ϕ_i should explicitly map subsets A_i of the acoustic space A into the BNF space B, rather than being determined by sub-classes Q_i of phones. This requires a better understanding of the topology of A and the relationships between its subsets and BPCs, which might be obtained through topological data analysis. In addition the ϕ_is should be homeomorphisms satisfying the consistency condition in Section 2. The latter could be investigated using 'reconstruction' DNNs in which the targets are equal to the inputs, although this might compromise the utility of the BNFs for classification.

Finally, experiments need to be conducted to confirm that the benefits of the local DNN structure for frame-level phone classification transfer to full ASR.

7. References

[1] L. Bai, P. Jančovič, M. Russell, and P. Weber, "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics," in *Proc. Interspeech,* Dresden, Germany, 2015, pp. 583–587.

[2] P. Weber, L. Bai, M. Russell, P. Jančovič, and S. Houghton, "Interpretation of low dimensional neural network bottleneck features in terms of human perception and production," in *Proc. Interspeech,* San Francisco, CA, USA, 2016, pp. 3384–3388.

[3] F. Li, A. Menon, and J. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2599–2610, 2010.

[4] F. Li, A. Trevino, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of american english fricative consonants in noise," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2663–2675, 2012.

[5] K. Stevens and S. Blumstein, "Invariant cues for place of articulation in stop consonants," *The Journal of the Acoustical Society of America*, vol. 64, no. 5, pp. 1358–1368, 1978.

[6] J. Heinz and K. Stevens, "On the properties of voiceless fricative consonants," *The Journal of the Acoustical Society of America*, vol. 33, no. 5, pp. 589–596, 1961.

[7] L. Raphael, "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in american english," *The Journal of the Acoustical Society of America*, vol. 51, no. 4B, pp. 1296–1303, 1972.

[8] L. Wilde, "Analysis and synthesis of fricative consonants," Ph.D. dissertation, Massachusetts Institute of Technology, 1995.

[9] A. Khasanova, J. Cole, and M. Hasegawa-Johnson, "Detecting articulatory compensation in acoustic data through linear regression modeling," in *Proc. Interspeech,* Singapore, 2014, pp. 925–929.

[10] P. Scanlon, D. Ellis, and R. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 15, no. 3, pp. 803–812, 2007.

[11] H. Huang, Y. Liu, L. ten Bosch, B. Cranena, and L. Boves, "Locally learning heterogeneous manifolds for phonetic classification," *Computer Speech and Language*, vol. 2016, pp. 28–45.

[12] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, 1993.

[13] J. Lee, *Introduction to topological manifolds*, 2nd ed., ser. Graduate texts in Mathematics 202. New York Dordrecht Heidelberg London: Springer, 2010.

[14] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[15] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," *Deep Learning and Unsupervised Feature Learning NIPS Workshop*, 2012.

[16] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. of the Python for Scientific Computing Conference (SciPy)*, June 2010.

[17] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE-ICASSP,* Glasgow, Scotland. IEEE, 1989, pp. 532–535.